文章からの感情等の特徴抽出について

平成 25 年 2 月 15 日

情報電子工学科 4 年
小椋 聡也

目 次

1 はじめに **1**

2 文章を解析について **1**

2.1 文章を解析する方法 1

2.2 構文解析について 2

2.3 文章中の文字数と感情の関係について 5

2.4 文章を解析するための指標について 6

3 実験 **6**

3.1 実験内容 6

3.2 実験に使用するプログラム 7

3.3 実験で解析を行う文章 7

3.4 平均や標準偏差等について 9

3.5 実験結果 11

4 考察 **18**

5 まとめ **23**

参考文献 **25**

概要

現在、電子メールは人々の間に広く普及しており、連絡等でメールが使われることが多く、その文章の中に相手の様々な感情が含まれていることが多い。人間であれば、文章を読むことで相手の感情を多少は理解することができるだろうが、文章中に含まれる感情等の特徴をコンピュータに自動的に判断させることができないかと考えた。電子メールを受信した時点でそのメール文章に含まれる感情を自動的に読みとることができれば、返信の際に、相手の感情を読みとれなかったことで相手に不愉快な思いをさせてしまうことを減らすことができるだろう。

過去の研究で、ブログの文章の解析を行い、感情等の特徴抽出を試みたものがあったので、そこで考察された、文章を解析する 3 つの方法をもとに、メールの文章から相手の感情等の情報を抽出する方法を考察した。そして、過去の研究で作成されていたプログラムを用いて実際に文章を解析し、過去の研究で今後の課題とされていた「文字数と平仮名の数は相関係数が高く、独立した指標とは言えないかもしれない」ということや他の指標間の関係に注目し、実験を行った。

過去の研究で解析された人物とは異なる 2 人のブログの文章を解析した結果、文字数と平仮名の字数の相関係数は高く、連動していると言えるため、平仮名の字数は独立した指標として使用することは難しいことがわかった。また、点数の計算方法は今回の計算方法では不正確だったため、より良い計算方法を考えることが今後の課題である。

1　はじめに

現在は、電子メールが広く普及しており、人との連絡やコミュニケーションにメールが使われることが多くなってきている。しかし、電子メール上の文章を見ても、メールの送り主がどのような感情でその文章を書いているのかわからないことがある。そのため、相手の文章に含まれた感情をよくわからずに返事を出して相手に不愉快な思いをさせてしまうこともあるかもしれない。しかし、メールを受信した時点でそのメール文章に含まれる感情をプログラムで自動的に判断することができれば、相手に不愉快な思いをさせることを少しでも減らすことができるのではないかと考えた。

過去の研究で、ブログの文章の解析を行い、感情等の特徴抽出を試みたものがあり, コンピュータでメール受信時に自動的に文章中の感情を判断してくれれば、多数の人からのメールが溜ってしまった際にどの人から優先的に返信すれば良いか等を考える時の判断材料にすることができると考えた。また、プログラムで文章中の感情を正確に判断することができれば、人間が見落としてしまった感情も見つけることができるだろう。

本研究では、過去の研究で考察された、3 つの文章を解析する方法をもとに、メールの文章から相手の感情等の情報を抽出する方法を考察する。

また、過去の研究で今後の課題として挙げられていた、「文字数と平仮名の数は相関係数が高く、独立した指標とは言えないかもしれない」とあったので、過去の研究で解析した文章とは異なる複数の著者の文章を解析し、他の人物でも同じことが言えるかどうか考察する。また、ブログの文章を人間が読み、その文章中に含まれる感情を判断したものとプログラムで文章を解析した結果をつき合わせ、実際にプログラムで感情を抜き出すことができるかどうかを検証する。

最終的に、受信したメールを自動的に解析し、評価するプログラムを作ることを目標とする。

2　文章を解析について

2.1　文章を解析する方法

ここで、過去の研究で挙げられていた文章を解析する 3 つの方法を紹介する。

(方法 1) 全文を読み込み、構文解析を行いその結果を元に解析する方法

> 全ての語句一つ一つに喜怒哀楽ごとに点数をあらかじめつけて、データベースに保存しておき、メールの文章を読み込んだ際にメールの文章に使われて

いる語句ごとに点数を呼び出し、その後、一文ごとに構文解析を行い、メールの文章全体の点数を出す。

構文解析については後述する。

(方法 2) 頭や語尾等の限られた一部だけに着目し、解析する方法

語頭や語尾等の一部の語句一つ一つに喜怒哀楽ごとに点数をあらかじめつけて、データベースに保存しておき、メールの文章を読み込んだ際にメールの文章に使われている語頭や語尾等の語句の点数を呼び出し、メール文章全体の点数を出す。この方法を用いる場合、構文解析は行わない。

しかし、この方法を用いるためには数多くある語句の一つ一つに喜怒哀楽ごとの点数をつけてデータベースに保存しなければならないため、かなりの手間がかかってしまう。

また、この喜怒哀楽ごとの点数は定められた基準は無く、作成者の判断で点数がつけられてしまうという問題点もある。

(方法 3) 語句の意味を調べず、文の数等を調べて解析する方法

この方法で調査する項目は文の数、字数、漢字の数等が挙げられる。語句の意味を全く調べないのであらかじめデータベースを作る必要もなく、構文解析の必要もない。

しかし、この方法では喜怒哀楽という感情をコンピュータに絶対的な評価をさせることは不可能なので、過去の文章との比較を行い、相対的な評価をさせる必要がある。

他にも文章を解析する良い方法がないか考えてみたが、あまり良い方法が思い浮かばなかった。また、方法 1 や方法 2 では単語ごとに喜怒哀楽の点数をつけて評価を行うため、単語データベース作成者によって結果が大きく左右されてしまう。よって、本研究では方法 3 の『語句の意味を調べず、文の数等を調べて解析する方法』を用いて、実験を行っていく。

2.2 構文解析について

構文解析とは、単語や字句で構成される文を、定義された文法に従って解釈し、文の構造を明確にすることである。

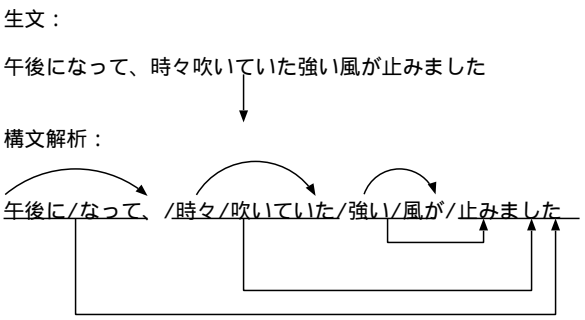

Fig. 1 構文解析の例

ここでは、構文解析を行う際に使うことが考えられる、kakasi と ChaSen という文章をわかち書きにすることができるソフトについての説明を行う。

- kakasi

kakasi は、漢字かな混じりの文章を平仮名文やローマ字文に変換することを目的として作られたプログラムと辞書のことである。また、わかち書き機能も実装されている。わかち書きとは、日本語の文章を語句毎に分割し、スペースで区切った状態のことを指す。

UNIX 上で kakasi を使用するときは以下のようにする。

- 標準入力から日本語文を入れる

kterm 上で以下のように入力する。

```
% echo ‘‘今日は雪なので、電車が遅れるだろう。’’ |kakasi -w
```

- 直接日本語を入力する

```
% kakasi -w
```

と kterm 上で入力したあと、

今日は雪なので、電車が遅れるだろう。

と入力する。終了するときは「Ctrl」+「D」キーを同時に押す。

上記のように入力し、実行すると出力は以下のようになる。

今日は 雪 なので 、 電車 が 遅れ るだろう 。

このように、kakasi を使えば簡単にわかち書きを行うことができる。

- ChaSen

ChaSen は、kakasi と同じように単語の区切りを調べるのが難しい日本語を単語毎に分割することができるソフトである。また、分割した単語の品詞の詳細を画面上に表示することができる。UNIX や Linux、MS-Windows で動作させることが可能。UNIX 上で ChaSen を使用する場合、以下のようにする。

- ファイルに日本語の文章を書き、そのファイルを ChaSen に読み込ませる

kterm 上で以下のように入力する。

```
% chasen file
```

- 標準入力から日本語の文章を入れる

kterm 上で以下のように入力する。

```
% echo ‘‘今日は雪なので、電車が遅れるだろう。’’ | chasen
```

- 直接日本語の文章を入力する

```
% chasen
```

と kterm 上で入力したあと、

今日は雪なので、電車が遅れるだろう。

のように入力する。終了する場合は、「Ctrl」+「D」キーを同時に押す。
上記のように入力し、実行した場合、出力結果は以下のようになる。

```
今日は雪なので、電車が遅れるだろう。
今日    キョウ  今日    名詞-副詞可能
は      ハ      は      助詞-係助詞
雪      ユキ    雪      名詞-一般
な      ナ      だ      助動詞  特殊・ダ        体言接続
ので    ノデ    ので    助詞-接続助詞
、      、      、      記号-読点
```

```
電車    デンシャ        電車    名詞-一般
が      ガ      が      助詞-格助詞-一般
遅れる  オクレル        遅れる  動詞-自立      一段    基本形
だろ    ダロ    だ      助動詞  特殊・ダ      未然形
う      ウ      う      助動詞  不変化型      基本形
。      。      。      記号-句点
EOS
```

kakasi は ChaSen に比べると軽いが、わかち書きの精度は ChaSen のほうが高い。また、kakasi のわかち書きでは文章を語句毎にうまく分けられない場合がある。例として、以下の文章を kakasi でわかち書きにすると、

この前から気になっていることがある。

上記の文章を kakasi でわかち書きにすると、

この 前か ら 気に なっていることがある 。

となる。この場合、” 前”、” から” となるべきなのだが、” 前か”、” ら” と分けられてしまう。一方、ChaSen では、

この 前 から 気 に なっ て いる こと が ある 。

となり、正しく区切られている。よって、構文解析を行うのであれば ChaSen のほうがふさわしいと考えられる。

2.3 文章中の文字数と感情の関係について

人間は、怒ったり喜んだりしていると文章中に以下のような特徴が現れるのではないかと考えた。

- 文の数が増えたり、字数も増えたりする
- 文章が短くなったり、字数が減ったりする
- !や... 等の記号が増える
- 極端に漢字や平仮名が増える

しかし、これらは文章を書いた人によって個人差があるため、その人が過去の文章と比較し、文章を書くときの癖や特徴等を調べ、文章にどのような傾向があるかどうかを調べて文章を解析して行く必要がある。また、漢字の数が多い文章を見ると堅苦しく、厳しいような印象を受け、漢字の数が少なく、平仮名が多い文章を見ると穏やかではあるが、少々幼稚な印象を受ける。このようなことも感情を判断する要素になり得るので、以上のことについても注目していく。

2.4 文章を解析するための指標について

以下に本研究で使用する文章を解析するための指標を示す。

- 本文相当行

  全体の行から、空行 (空けてある行)、引用行 (返信の場合そのまま受信したものを引用してあるもの、> の記号で表されることが多い)、署名と署名に相当する行 (名前、メールアドレス、電話番号など) を引いたもの。

- 全文字数

  本文中の全ての文字の数。

- 一文の平均字数

  全文字数÷ 本文相当行で求められる。一文の平均字数。

- 区切文字数

  句読点等 (句読点と記号類!、?、「 」なども含む) で文章が区切られていると考えられる部分の文字数をそれぞれ数え、区切の数で割った、区切り内の文字数平均。

- 区切数

  全区切数÷ 本文で求められる。一文につきいくつの区切りがあるかの平均値。

3 実験

3.1 実験内容

本研究の最終的な目標は、文章から相手の感情を自動的に判断し、評価するプログラムの作成だが、ここでは過去の研究で作られた文章を解析するプログラムを用いて、実際に文章を解析し、平均や標準偏差などの結果から指標や文章の特徴について考察を行う。以下に実験の手順を示す。

1. 特定の人間が書いた文章を複数、プログラムに入力する

   プログラムは Perl というプログラム言語を用いる。Perl については 3.2 節で説明する。

2. それぞれの文章毎に各指標の値を出力させる

   指標については 2.3 節を参照。

3. 出力された結果から平均や標準偏差等を計算し、出力する

   平均や標準偏差、相関係数、点数を求めるための計算式を 3.3 節に示す。その式を用いて平均や標準偏差、相関係数、点数の値を求め、出力する。

4. 計算結果を元に考察する

   平均や標準偏差、相関係数、点数の値を見て、文章の特徴や指標間の相関等について考察する。

3.2 実験に使用するプログラム

本実験では、過去の研究で作成された文章を解析するプログラムを用いる。このプログラムは、Perl というプログラミング言語を用いて作成されている。

Perl はラリー・ウォール (Larry Wall) 氏が開発したプログラミング言語であり、かつてはスクリプト言語と呼ばれることもあったが、現在では両者はあまり区別されないのが実体である。

Perl の言語文法はとにかく記号を多用するため、慣れないと少々使いづらいだろう。しかし、コンパイルと呼ばれる機械語に変換する作業が必要無く、プログラムがテキスト形式なので作成や実行、修正が容易にできる。また、C、awk、sed、シェルスクリプト等のほとんど全ての機能を取り込んでいるため、それらの言語でできることで Perl にできないことはほとんど無いが、ひとつのことを実現するのに何通りでもできてしまうといった「副作用」がある。また、Perl は C 言語等のプログラミング言語と比べるとテキスト処理が強力であり、正規表現を演算子として使用できることも大きな長所である。そのため、Perl は世界中のプログラマに愛用されている。

3.3 実験で解析を行う文章

本実験では、文章を読み込んで解析を行うので、サンプルとなる文章が必要である。また、サンプルとして使用する文章はある程度の長文であることが望ましい。なぜなら、本

実験では文章中の全ての文字数を指標としているため、短文をサンプルとして使用した場合、全文字数に大きな違いがなくても結果が大幅に変わってしまうことがあるからである。また、本実験では顔文字や絵文字を解析するのは難しいため考えないものとする。

以上のことから、ある程度長文であり、絵文字や顔文字がほとんど使用されていない品川ブログと室井佑月ブログというブログの文章をサンプルとして使用する。例として、以下に品川ブログ [4] の文章を載せる。

2012 年 10 月 31 日（水）
チャラチャラ
今日は朝から仕事でした。
それも朝っぽくない芸人。
オリラジの藤森。
カラテカの入江。
チャラ男
アンド
チャラおじさん
チャラ族はハットにメガネがマストらしいな！
仕事を終えてチャラチャラコンビとランチ。
にしても、
彼らの噂話好きには驚いた。
芸能界の噂
吉本の内部事情
六本木西麻布の都市伝説
まるで漫オコンビのように、交互に、テンポよく、情報を放り投げてきやがる。
東京入江スポーツと
チャライデー。
恐ろしいやつらだ。
でも、
僕が聞く限りガセネタも混ざっていたので...
聞いて楽しむ程度の情報ですな～
ちなみに僕はその後、招待された試写会へ...
チャラチャラコンビは宝塚へ
ベルばらを見るらしい。
宝塚歌劇団のみなさん。
ゴシップ大好きな二人が潜入しました。
お気をつけくだされ！

上記の文章をプログラムに入力し、解析した結果を以下に示す。

文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
28	353	87	132	80	20

Table 1: 品川ブログの文章を解析した結果

解析結果は以上のようになり、文章の点数は -0.0440 点だった。点数の計算方法については次節で説明する。このように、ひとつひとつ文章を入力して文章を解析し、実験を行う。

3.4 平均や標準偏差等について

- 平均

  変数を x_1,x_2,...,x_n, その項数を N とすると、算術平均 $\bar{x}$ は次式で定義される。

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

  平均値は度数がどのような値を中心にして分布しているかという意味で、度数分布の位置を示す測定値である。分布の中心の所在を示す値であるが、必ずしも分布の中央、あるいは度数の正確な集中点とは一致しない。平均値の計算の仕方と分布の形によっては、平均値が分布の中央または集中点に一致する場合もあるが、一般にはもっと広い意味での分布の中心であって、変数の全体をひとまとめにして代表する値である。

- 標準偏差

  平均値に対する変数の偏差を平均するとき、正負の符号を処理する方法として、偏差の絶対値をとるかわりにその 2 乗値を使用するのが標準偏差および分散である。平均値に対する偏差を 2 乗したものを平均し、これを変数と同じ次元の量で示すために平方に開いたものを標準偏差といい、s で表す。

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2}$$

- 分散

  分散 v とは、データの散らばり具合の大きさを表す値である。また、この分散の平方根をとった値のことを標準偏差という。

$$v = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{3}$$

- 不偏標準偏差

  過去の研究で作られたプログラムでは標準偏差ではなく、不偏標準偏差を用いて計算を行っていたので、その不偏標準偏差についても説明する。

  母集団からデータの標本をとって計算する場合、その標本の分散の値は平均して母集団の分散とは少しずれが生じることが知られている。そのずれを無くすため、平均が母集団の平均になるように調整したものが不偏分散 v_1 であり、その平方根をとったものを不偏標準偏差 s_1 という。

$$v_1 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{4}$$

$$s_1 = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{5}$$

- 相関係数

  2 変数 x,y の関係が単純な相互依存の関係のとき、相関分析の方法で両変数間の相互依存関係が計測され、関係の方向と強さを示す測度として相関係数が使用される。相関係数の数値が 1 に近い程関係が強く、連動して変化しているため、独立した指標とは言えないことになるだろう。相関係数が 0 に近い場合、お互いの相関が弱いことを意味している。相関係数は次式で定義される。

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{6}$$

相関係数 r の値の変域は $-1 \leq r \leq +1$ であることが示される。

$r > 0$ のときを正の相関または順相関、$r < 0$ のときを負の相関または逆相関という。$r = \pm 1$ のときを完全相関、$r = 0$ のときを無相関という。

相関係数 r は、x, y の間の関係がどの程度直線的であるかを示す測度である。一般に 2 変数 x, y が互いに無関係であれば $r = 0$ となるが、$r = 0$ であっても、必ずしも無関係とは限らない。

- 点数

この「点数」は文章中の感情を点数化したものではなく、ある 1 人の人物が今まで書いてきた文章の各指標毎の平均値と今回の各指標毎の値がどれだけ異なっているかを点数化したものである。よって、点数が 0 に近くなるほど、各指標の値が平均値に近い「いつも通りの文章」と言え、点数が 0 から遠くなるほど、各指標の値が平均値と大きく異なる「いつもとは違った文章」と言える。

ここでの「点数」は、ある 1 人の人物の過去の文章を解析した結果から各指標毎の平均値を求め、その各指標毎の平均値と今回の各指標の値との差をその指標の最大値で割り、それらの和をとって計算した。

点数 p は以下の式で定義される。本実験では、指標数は 6 つなので $N = 6$ であり、$\bar{x}_i$ は今までの指標 i の平均値、m_i は今までの指標データの最大値、x_i は今回の指標 i の値を表している。

$$p = \sum_{i=1}^{N} \frac{x_i - \bar{x}_i}{m_i} \tag{7}$$

3.5 実験結果

本実験では 11ヵ月分のブログの文章をサンプルとして解析を行った。今回の実験で品川ブログの文章を解析し、得られた指標データを Table 2 に示す。また、比較対象として室井佑月氏の 11ヵ月分のブログの文章も解析したので、その結果を Table 4 に示す。

文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
28	353	87	132	80	20
14	297	78	138	27	30
20	380	76	211	22	32

文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
28	405	108	214	14	28
92	1326	280	694	123	120
10	154	39	84	9	9
49	672	118	328	93	51
26	318	63	164	34	19
17	276	60	144	18	17
41	568	105	294	80	41
49	490	130	213	52	35
60	731	149	363	72	58
49	906	207	479	87	48
47	563	105	313	41	52
11	160	22	88	25	13
55	786	177	455	42	50
27	374	58	157	59	19
108	1325	279	688	110	88
54	648	143	337	66	35
134	1944	369	1085	115	183
42	471	119	214	39	46
40	582	98	222	149	31
41	486	77	268	58	31
25	287	69	148	19	19
17	245	35	122	49	16
20	237	63	104	30	18
38	486	94	236	58	49
28	671	108	376	102	52
30	428	78	223	57	31
14	183	22	89	43	13
9	102	18	40	22	9
40	564	87	313	64	45
41	575	98	322	53	39
37	477	94	230	51	44
10	133	31	70	3	7
8	132	24	69	19	10
7	185	41	80	31	7
15	168	50	86	5	10
5	62	17	28	10	3
10	203	59	95	29	10

文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
17	159	38	78	20	3
30	365	87	200	24	23
12	165	41	88	17	6
22	209	60	105	3	13
11	106	16	46	22	11
26	431	76	194	97	23
24	308	48	150	66	16
6	46	21	15	3	2
41	478	111	241	31	29
23	373	83	211	20	23
5	103	12	45	27	9
10	150	30	87	14	8
17	231	31	123	33	11
32	399	66	194	60	27
19	293	78	123	48	19
55	833	154	460	104	45
14	234	54	134	12	18
36	463	100	229	59	28
10	156	35	66	21	11
17	56	55	51	13	15
37	542	104	319	28	50
23	268	78	108	30	21
12	164	25	64	46	11
34	647	87	293	160	29

Table 2: 品川ブログ文章を解析したことで得られた指標データ

	文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
文の数	-	0.969	0.956	0.955	0.679	0.930
全ての字数	-	-	0.970	0.992	0.746	0.948
漢字数	-	-	-	0.956	0.645	0.916
平仮名の字数	-	-	-	-	0.689	0.951
片仮名の字数	-	-	-	-	-	0.616
区切り字数	-	-	-	-	-	-

	文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数

Table 3: 各指標間の相関係数

Table 3 に各指標間の相関係数について示す。この表をみてわかることは、文章中の全ての字数と平仮名の字数の相関係数が 0.992 と 1 に非常に近い値になっているため、文章中の全ての字数が多くなるほど平仮名の字数も連動して増えていくと言える。また、片仮名の字数以外の指標の相関係数も全て 0.9 を上回る値になっているため、それぞれの指標間に強い関係があることがわかる。

次に室井佑月ブログ [5] の文章と文章を解析した結果を示す。

< 飲むぞーっ！ >

５月１日　午後一時
みんな、ゴールデンウィーク、楽しんでいるかい？　なんでも今年のゴールデンウィークはここ数年の中で、いちばん海外に出る人が多いんだとか。
ま、あたしには関係ない話だな。
原稿の締め切りが９日に集中しているでやんの。編集様が長期休暇を取るために。
くそー。尻が痛い。もう何時間机に向かっているんだか。
７日の飲み会だけが楽しみじゃ。
飲み会、このブログに書いたあたしの友達いっぱい呼んだから。
ぎゃあぎゃあみんなで話をする楽しい会にしようね。
飲むぞ、飲むぞ、飲むぞーっ！

文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
12	239	52	130	17	19

Table 4: 室井佑月ブログの文章を解析した結果

文章の点数は 2.123 点であった。

文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
53	1351	296	787	77	114
66	1463	244	921	86	133
61	1418	280	889	47	118
33	705	161	423	20	55
35	638	128	383	21	50
47	1065	196	642	66	89
35	742	168	416	53	59
33	690	194	360	25	54
42	1057	202	633	50	99
22	447	98	257	20	33
66	1259	245	711	75	96
68	1324	249	752	111	128
38	905	178	582	27	64
55	1048	194	658	41	95
28	665	127	411	38	48
80	1951	350	1270	88	154
12	239	52	130	17	19
43	868	199	470	44	91
87	2249	543	1281	96	186
85	1890	314	1264	45	169
64	1403	310	796	87	123
117	2769	575	1790	47	240
52	1256	226	785	89	104
31	638	128	382	35	57
49	1020	211	543	105	94
49	1020	211	543	105	94
63	1579	322	916	108	134
33	929	205	553	37	83
75	1613	336	996	52	133
37	825	151	535	19	73

Table 5: 室井佑月ブログの文章を解析したことで得られた指標データ

	文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
文の数	-	0.976	0.924	0.967	0.500	0.974
全ての字数	-	-	0.965	0.991	0.492	0.988
漢字数	-	-	-	0.934	0.460	0.947
平仮名の字数	-	-	-	-	0.415	0.976
片仮名の字数	-	-	-	-	-	0.520
区切り字数	-	-	-	-	-	-

Table 6: 各指標間の相関係数

Tabular 5 に室井佑月ブログの各指標間の相関係数について示す。この表をみてわかることは、品川ブログと同じように文章中の全ての字数と平仮名の字数の相関係数が 0.991 で 1 に近い値になっていることである。また、片仮名の字数以外の指標の相関係数も品川ブログと似たような値になっていた。

さらに、今回の実験結果と比較するために、過去の研究で解析されていた文章の例とその実験結果を以下に示す。なお、過去の研究では MORI LOG ACADEMY [3] というブログの文章をサンプルとして文章の解析を行っていた。

【HR】 ひさしぶりの日記

子供の頃にも日記はつけて続いたためしはなかったし、大人になってもそういった習慣はないのに、インターネットのサイトでかつて日記を始めたら、これが 5 年も続いて 5 冊の本になった。今まで書いてきたどの著作よりも、その日記が自分では一番の力作だと思っている。最初から、出版するつもりでかいたから、つまり仕事としてやったから続いたと思う。でも、マンネリを感じて 2001 年でいったんやめてしまって、そのときは「ああ、もう書かなくて良いのだ」と本当に嬉しかった。きっと、小説をやめたら、同じくらい嬉しいだろう。

さて、日記をもう一度書いてみることにした。今回も仕事として書くし、出版物になる予定だから、続くだろう。前回は、「ネットの日記を本にするなんて」という目で見られたけれど、今では普通になった。もの凄い大勢の人がネットで日記を書いている。いったい読み手はいるのか、と不思議。そういう話はまたそのうちゆっくりと … 。

文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
62	1601	354	910	83	147
55	1461	350	791	79	140
37	999	216	560	51	119
40	1149	274	625	102	107
49	1199	311	642	81	118
55	1562	376	847	108	149
45	1231	318	592	124	111
42	1106	246	654	33	137
56	1167	419	910	99	152
50	1196	320	593	114	98
44	1135	320	595	114	98
55	1516	315	877	106	169
57	1484	385	829	81	136
39	1078	271	600	27	99
59	1824	450	1095	82	143
63	1716	448	965	105	128
61	1636	409	844	109	158
55	1303	358	705	45	114
39	1135	237	669	61	121
56	1495	407	846	60	132
50	1432	414	801	50	121
55	1479	360	811	69	144
45	1505	361	899	50	150
59	1439	354	728	170	129
37	1038	247	604	47	103
53	1557	399	919	37	140
53	1708	440	1013	89	128
32	1039	234	617	68	82
40	975	219	534	93	90
50	1424	320	769	84	179

Table 7: 過去の研究で解析された指標データ

	文の数	全ての字数	漢字数	平仮名の字数	片仮名の字数	区切り字数
文の数	-	0.850	0.764	0.805	0.364	0.723
全ての字数	-	-	0.888	0.955	0.432	0.867
漢字数	-	-	-	0.815	0.217	0.713
平仮名の字数	-	-	-	-	0.264	0.810
片仮名の字数	-	-	-	-	-	0.348
区切り字数	-	-	-	-	-	-

Table 8: 過去の研究で求められた各指標間の相関係数

過去の研究で解析された MORI LOG ACADEMY の各指標間の相関係数は、他の 2 人と比べると全体的に値が低かった。また、もっとも相関係数が高かったのは “全ての字数” と” 平仮名の字数” の組合せだった。

4 考察

実験によって得られた指標データの関係をグラフ化したものを以下の Fig2 ~ 9 に示す。

本研究では 2 人の文章を解析し、“全ての字数” と” 平仮名の字数” の相関係数を調べたが、品川ブログの “全ての字数” と” 平仮名の字数” の相関係数の値は 0.992、室井佑月ブログの “全ての字数” と” 平仮名の字数” の相関係数の値は 0.991 と両者とも 1 に近い値になっていた。また、過去の研究で求められていた “全ての字数” と” 平仮名の字数” の値も 0.955 で 1 に近い値になっていた。また、Fig.2,3,4 を見るとグラフが重なり合っている部分が多いため、2 つの指標が連動していることがわかる。そのため、“全ての字数” と” 平仮名の字数” はそれぞれが独立した指標とは言えないことになるだろう。よって、この” 平仮名の字数” という指標は不要であると考えられる。また、Fig.5,6,7 を見ると、” 漢字数” と” 片仮名の字数” のグラフが重なり合っている部分が少なく、相関が低いことがわかる。そのため、” 漢字数” と” 片仮名の字数” はそれぞれが独立した指標として使用することができると考えられる。3.5 節で示した、室井佑月ブログの文章は人間が見ても喜んでいることが読み取れる文章であり、プログラムによる評価も 2.123 点であったため、平均的な文章とは大きく異なる文章であると言える。3.5 節で示した文章では人間の評価とプログラムの評価が一致したが、他の文章ではそうではないものが多く存在した。そのため、この評価方法では正確に文章中の感情を判断することは難しいと言うことが分かった。また、この点数は、平均的な文章との違いを点数化したものなので、点数を見るだけでは怒っているのか、喜んでいるのか判断することも難しい。これらの点は今後の課題である。

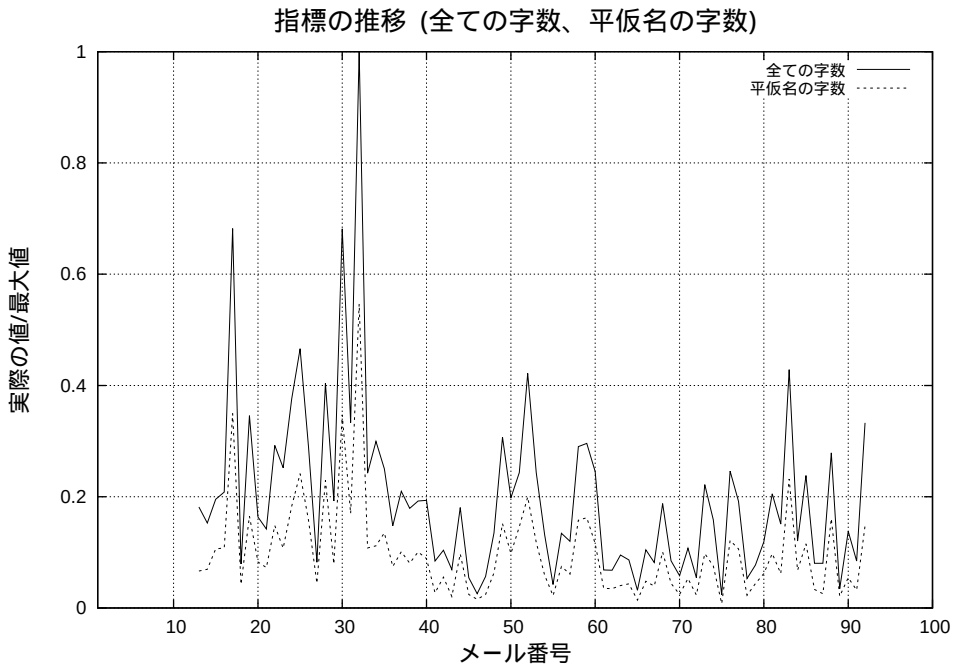

Fig. 2 品川ブログの全ての字数と平仮名の字数の関係

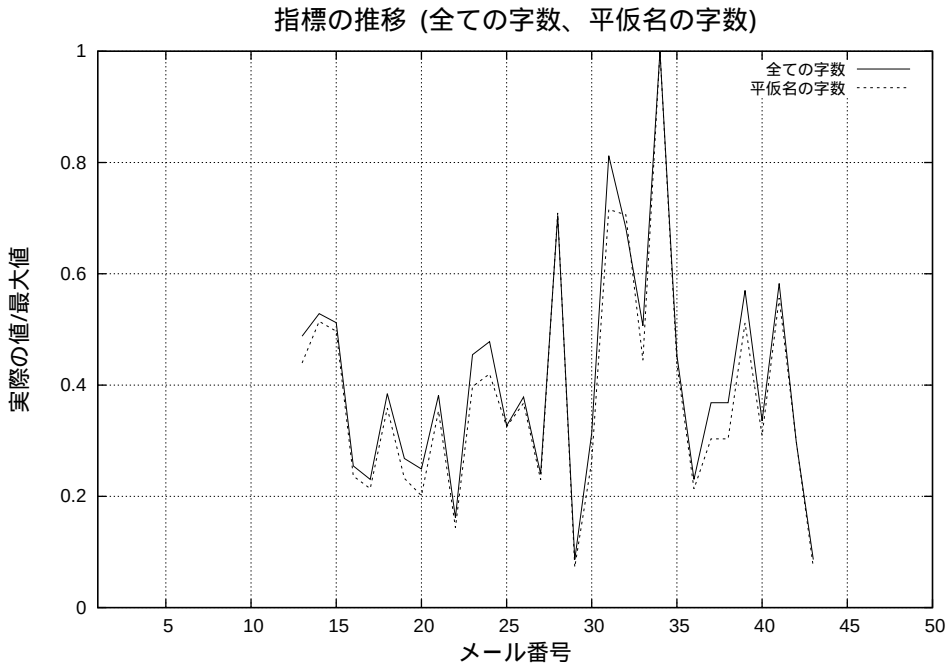

Fig. 3 室井佑月ブログの全ての字数と平仮名の字数の関係

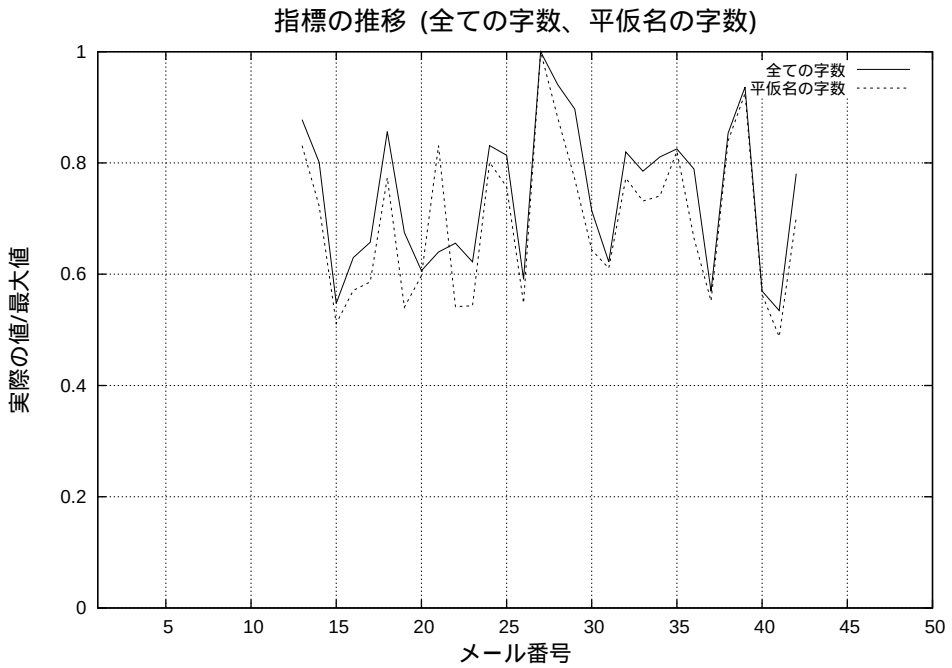

Fig. 4 MORI LOG ACADEMY の全ての字数と平仮名の字数の関係

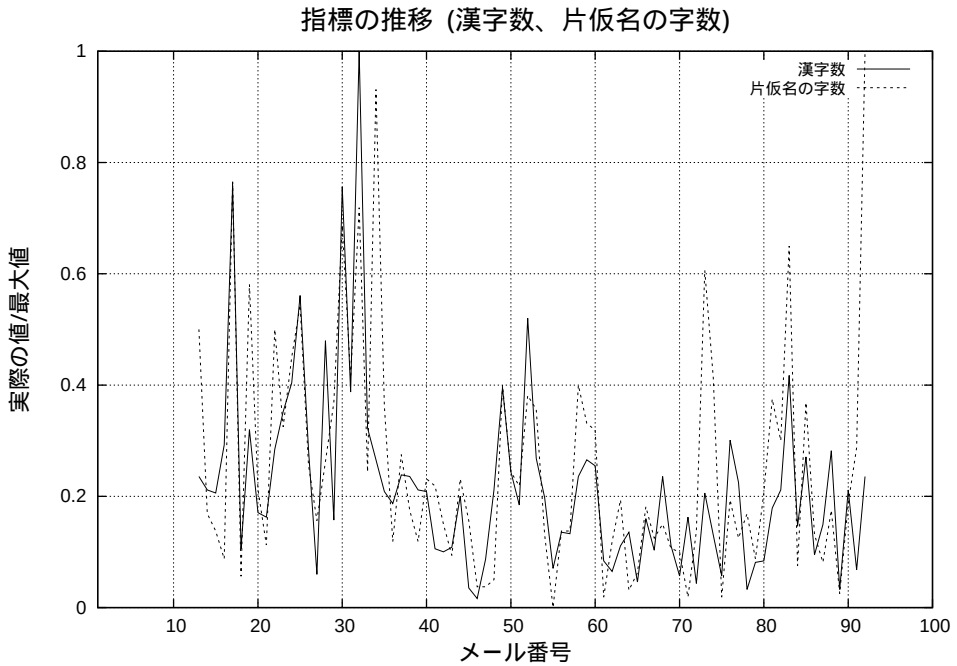

Fig. 5 品川ブログの漢字数と片仮名の字数の関係

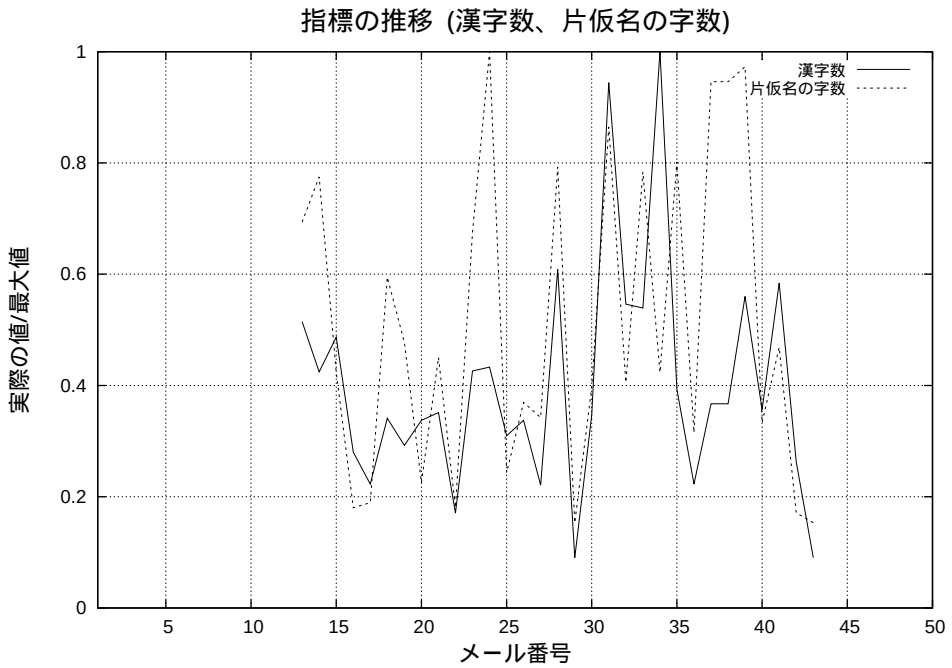

Fig. 6 室井佑月ブログの漢字数と片仮名の字数の関係

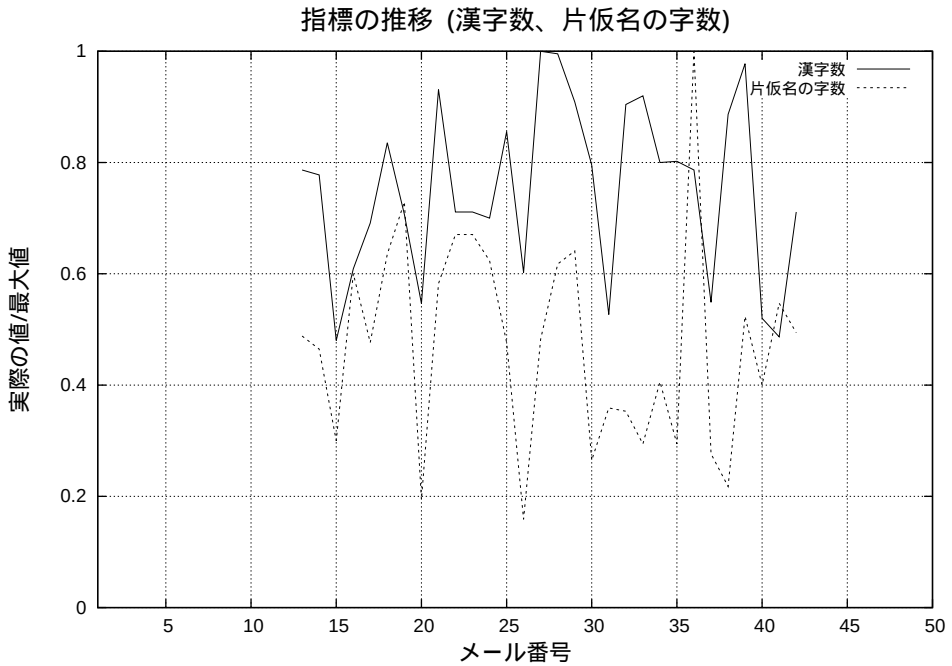

Fig. 7 MORI LOG ACADEMY の漢字数と片仮名の字数の関係

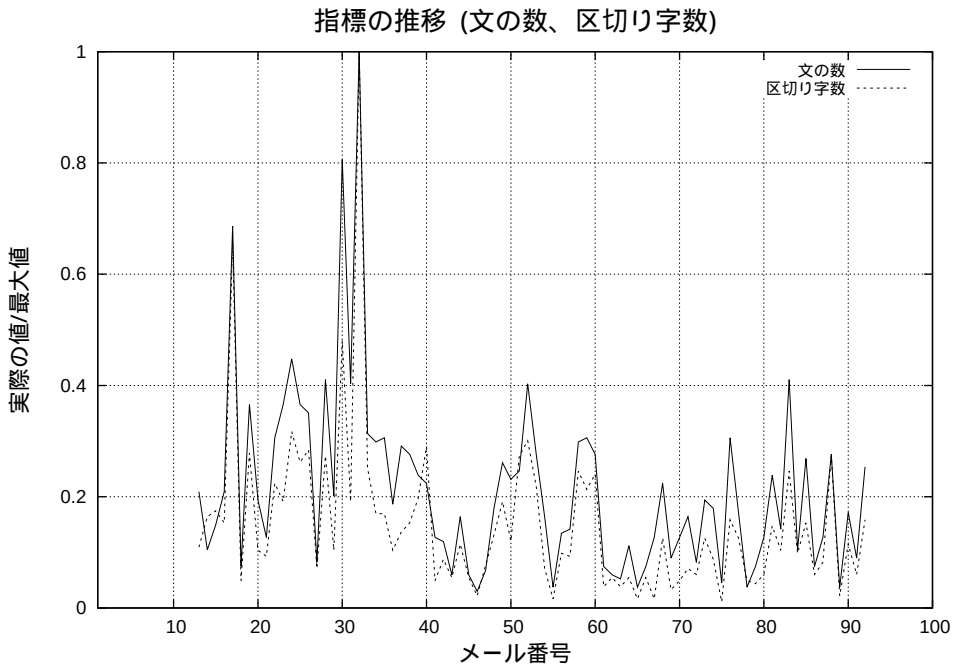

Fig. 8 品川ブログの文の数と区切り字数の関係

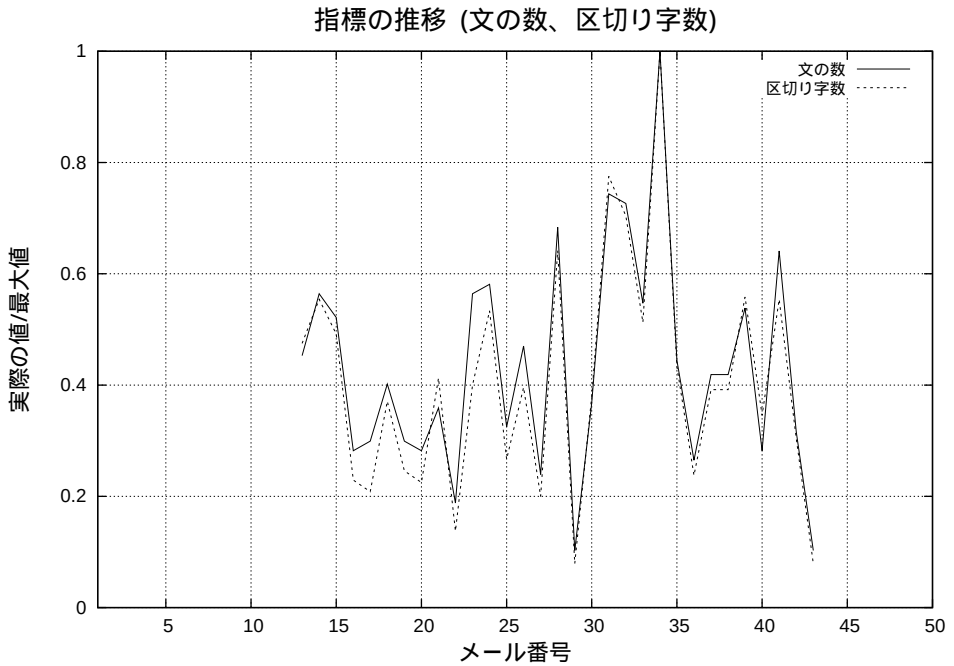

Fig. 9 室井佑月ブログの文の数と区切り字数の関係

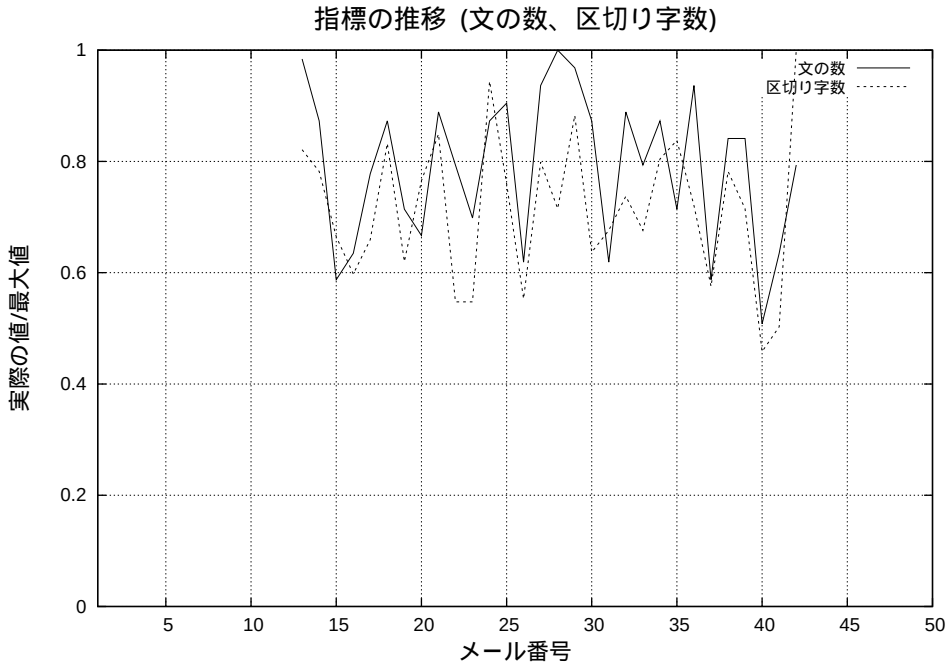

Fig. 10 MORI LOG ACADEMY の文の数と区切り字数の関係

5 まとめ

電子メールを受信した時点でそのメール文章に含まれる感情を自動的に読みとることができれば、返信の際に、相手の感情を読みとれなかったことで相手に不愉快な思いをさせてしまうことを減らすことができるかもしれないと考えた。また、コンピュータでメール受信時に自動的に文章中の感情を判断してくれれば、多数の人からのメールが溜ってしまった際にどの人から優先的に返信すれば良いか等を考える時の判断材料にすることができるだろう。本研究の最終的な目標は、メールの文章を自動的に読み込んで文章中から相手の喜怒哀楽等の感情を点数化して評価することであるが、ここでは過去の研究で考察されていた、文章を解析する 3 つの方法や文章を解析するプログラムを用いて文章を解析する実験を行い、指標について比較、考察を行った。また、過去の研究で今後の課題とされていた、「“全ての字数” と” 平仮名の字数” は独立した指標か」や「ブログの文章中の感情を人間が判断し、プログラムによって求められた結果とつき合わせる」ということについても考察を行った。そして、文章を書いた人物が異なると相関係数などにどのような違いが生じるか考察した。

そして今回、過去の研究と合わせて 3 人の文章を解析したが、3 人とも” 全ての字数” と” 平仮名の字数” の相関係数の値が高かった。過去の研究で今後の課題とされたものの中に「“全ての字数” と” 平仮名の字数” は独立した指標と言えるかどうか」とあったが、今回 3 人の文章を解析した結果から、” 平仮名の字数” は独立した指標と言うことは難し

いことがわかった。しかし、” 漢字数” と” 片仮名の字数” は 3 人とも他の指標間の相関係数よりも低い値になっていた。よって、これらの指標は独立した指標として使用することができるだろう。

「ブログの文章中の感情を人間が判断し、プログラムによって求められた結果とつき合わせる」については、3.5 節の室井佑月ブログのサンプル文章を解析した結果では 2.123 点という高得点であり、人間がその文章を見ても楽しそうだということが読みとれるため、うまく評価できたといえるだろう。しかし、人間が見たら楽しそうに見える文章でもプログラムの点数が低いということも多く見られた。そのため、この評価方法ではあまり良い評価ができないと言うことがわかった。そのため、新たにより正確な評価方法を考えることも今後の課題である。

参考文献

[1] 遠藤建城: メール文書からの感情等の特徴抽出について, 新潟工科大学工学部情報電子工学科卒業論文 (2009 年)

[2] 森田優三、久次智雄: 新統計概論, 日本評論社 (1993 年)

[3] 森博嗣: MORI LOG ACADEMY1 ~ 3 (メディアファクトリー,2006)

[4] 品川ヒロシ: 品川ブログ (ワニブックス,2007)

[5] 室井佑月: 室井佑月ブログ, http://muroi-yuzuki.cocolog-nifty.com/